DOCUMENT RESUME

ED 259 006 TM 850 375

AUTHOR Perlman, Carole L.; And Others
TITLE Choosing an Achievement Battery: A Case Study.
PUB DATE Apr 84
NOTE 13p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 21-23, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Descriptive (141)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Achievement; *Achievement Tests; Criterion Referenced Tests; Elementary Secondary Education; *Evaluation Criteria; *Grade Equivalent Scores; Measurement Objectives; Political Influences; Reading Achievement; *Reading Tests; Standardized Tests; *Test Selection
IDENTIFIERS California Achievement Tests; Iowa Tests Of Basic Skills; Metropolitan Achievement Tests; Sequential Tests Of Educational Progress; Sra Achievement Series

ABSTRACT

This paper provides school personnel with a practical outline for selecting an appropriate standardized achievement battery by describing the test selection procedures developed by a large metropolitan school system. The system needed a new achievement test battery which would reflect recent curriculum changes and provide for grade level testing. Evaluation concerns included norms, reliability, scaling, content, format, costs, item bias, and accountability to the public. During the first phase of the selection process, 2,000 students in grades 1-8 took each of five standardized reading achievement tests. Reliability, item difficulty, percent of students scoring at the floor or ceiling of each test, and the relationship between raw scores and grade-equivalent scores for each test were examined. After the first phase, two tests remained. The second phase emphasized how well the tests measured what was being taught, especially reading comprehension. The old test and both suggested new tests were administered to 35,000 students in 100 schools. Information on students' reading progress was studied in relation to grade equivalent scores. The final selected test had moderate length and difficulty, good reliability, and low percentage of scores either below the chance level or above 80 percent. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED259006

CHOOSING AN ACHIEVEMENT BATTERY:

A CASE STUDY

Carole L. Perlman

Linda K. Junker

William K. Rice, Jr.

Chicago Public Schools

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Perlman, Carole

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 850 315

CHOOSING AN ACHIEVEMENT BATTERY: A CASE STUDY

Carole L. Perlman, Linda K. Junker, William K. Rice, Jr.
Chicago Public Schools

Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, 1984.

Objective

Although the testing literature contains a number of guidelines on how to select tests (e.g., American Psychological Association, 1974), there are few studies on the actual implementation of those guidelines. The purpose of this paper is to provide school personnel with a practical outline for selecting an appropriate standardized achievement battery. Test content, technical quality, and other issues are considered. An actual case study is given in which the procedures described are used.

Background

The procedures described in this paper were developed during a two-year test selection process in a large metropolitan school system, but they could be modified and implemented in a much shorter time and in a much smaller school system. The system had been using its current test battery for seven years and the norms were outdated. Functional level testing was being used, but there was some sentiment in favor of a return to grade level testing. There was concern that the test did not adequately reflect recent changes in curriculum. Clearly, it was time to find a new test.

Responsibility for selecting the new test was divided between two groups. Their orientations could basically be classified as either technical (interested mainly in norms, reliability, scaling) or curricular (primarily concerned with the test content). Other test selection criteria included test format, cost, and freedom from item bias. It was necessary to develop selection procedures that would adequately reflect the interests of both groups and provide the school system with a test that would best meet the needs of evaluation, curriculum development, and accountability to

the public. One test selection procedure that seemed promising was a study in which several nationally normed achievement batteries would be pilot tested on comparable groups of local students.

Instruments Used

The tests that were included in pilot tests were the Iowa Tests of Basic Skills, Form 7; the California Achievement Tests, Form C; the Sequential Tests of Educational Progress, Form X; the SRA Achievement Series, Form 1; and the Metropolitan Achievement Tests, Form JS.

Method

The test selection proceeded in two phases. In the first phase, a list of desirable properties for an achievement test was developed (see Figure 1).

Insert Figure 1 about here

Elementary and high school committees were convened to study the tests and pilot testing was conducted with the five tests listed above. Students were tested twice, once at their functional reading level and then randomly assigned to either one level above or one level below their functional level. Approximately 2000 students in grades 1-8 took each test. The following variables were empirically determined for each level of each test: reliability, item difficulty, and the percentage of students scoring at the floor or ceiling of each test. The relationship between raw scores and grade-equivalent scores for each test was examined. (The school system was committed to the use of grade-equivalents.)

There was considerable concern about whether the tests measured what was being taught, since some people felt that the reason that scores on the old test were low was because that test did not adequately correspond to the

Figure 1

Criteria for Selection of an Achievement Battery

In general, the test should conform to the American Psychological Psychological Association's Standards for Educational and Psychological Tests in the areas of dissemination of information; interpretation; validity; reliability; administration and scoring; and scales and norms. Some relevant points to consider are given below.

Test reliability: Is the test internally consistent (reliability of at least .85)? Is a pupil's score measured accurately and is it stable over a short time interval (test-retest reliability of at least .80)?

Content validity: Do the test items actually constitute a sample of the tasks they are claimed to represent?

Applicability: Is the test content consistent with the local curriculum, especially in reading and mathematics? Are test objectives available and do the test items really measure those objectives?

Construct validity: Do the tests measure the attributes they are said to measure? Do they measure extraneous things? For example, to what extent is a mathematics test actually measuring the student's ability to read word problems?

Freedom from racial, cultural and sexual stereotyping.

Adequate norms: Has the test been normed on an appropriate population including students from large cities? Has the test been normed recently? Are empirical norms available for several points throughout the school year? Is the test normed in such a way that there is a single continuum of growth from each test level to the next?

Adequate manual: Does the test manual provide sufficient reliability and validity data? Are the test instructions clear? How good are the instructions for interpreting test scores?

Availability of parallel forms.

Ease of implementing functional level testing in both reading and mathematics.

Types of scores provided: Are standard scores, normal-curve-equivalents, percentile ranks, stanines and grade-equivalent scores available? Is any special diagnostic information given?

Comprehensiveness: What subtests outside of mathematics and reading does the battery contain?

Appropriateness of the test for pupil placement and program evaluation.

Cost: How expensive is the test, both initially and on a long-term basis? Can it be scored on existing in-house optical scanning equipment?

Range: Are tests available for students from kindergarten through high school?

Length: Is the test long enough to provide reliable measurement, but not so long as to be cumbersome?

local curriculum. Some curricular studies were done, but not in a systematic fashion. Based on technical characteristics, curriculum studies, and presentations from the test publishers, the test selection committees were still unable to select a single test, although they were able to narrow the field to two tests (the elementary committee favored one test and the high school committee another). One test seemed to be preferable on technical grounds and the other on the agreement of the test with local curriculum.

After the field was narrowed to two tests, the second phase of the test selection process began with more pilot testing and curriculum studies. The problem was to determine how well the tests measured what was being taught in the local schools, with special emphasis on reading comprehension.

The school system's reading curriculum has a set of units at each grade and a set of criterion-referenced tests to measure mastery of each unit. It was decided that the fairest way to compare the two tests was to perform a blind curriculum matching study. Items were retyped and renumbered in such a way that raters could not determine which test or test level they originally came from. The items were divided into two sets, primary and intermediate-upper. Three raters, who were experienced reading resource teachers, were assigned to each set. Each rater was asked to determine at what grade and what unit the content of each test item was first taught, and at what grade(s) they would expect 50% and 80% of the students to get the item correct.

One hundred schools with approximately 35,000 students were randomly selected to participate in the second pilot study. Schools were randomly assigned to administer both the currently used test and one of the two new tests (the order of administration was counterbalanced). In addition, information was obtained on the students' progress in the reading curriculum, specifically what grade level of reading materials they were working on and how many of the units at that level that they had mastered. The variables of interest were the grade-equivalent scores obtained and the relationship between those scores and the students' placement in the reading curriculum.

Results

During the first phase of test selection, the test that was ultimately chosen had the following technical characteristics: subtests that were moderately long to long; items with moderate difficulty; a relatively low percentage of students scoring below the chance level or above 80% correct; a low percentage of items where a raw score change of one point results in a grade-equivalent score change of half a year or more; a low maximum possible grade-equivalent change caused by a one-point change in raw score; and good reliability. (A summary table which gives exact figures for all five tests is presented in Table 1).

Insert Table I about here

During the second phase of test selection, the blind curriculum match indicated that both of the remaining tests covered about the same number of units in the local reading curriculum. This was contrary to the findings of the previous, less structured, curriculum matching studies that had shown one test to be clearly more in line with local curriculum than the other. Furthermore, scores on the test whose content had initially been thought to be more "relevant" (from the point of view of the curriculum experts), actually showed worse correspondence with students' reading curriculum placement that did the other test. Interestingly enough, the test that was rejected was ruled out partly because, at the upper grades, the grade-equivalent scores actually looked too high (based on the students' reading curriculum placement and previous test scores). It was felt that a sudden, sharp increase in test scores at the upper grades would be difficult to explain to the public. Indeed, an editorial criticizing use of the "easier" test appeared in a local newspaper (Chicago Sun-Times, 1982).

BEST COPY AVAILABLE

Table I

Summary of Technical Results

	CAT	ITBS	Metro	SRA	STEP/CIRCUS
st length	20 items at lowest two levels; 27 at next level; 40 at remaining levels. Total # of levels is 8.	44-67 items; all but two levels have over 50 items. Total # of levels is 9.	55 or 60 items at each level. Total number of levels is 5.	First four levels have 35, 24, 24, 26 items, respectively; other levels have 50 items. Total number of levels is 7.	42 and 45 items at lowest two levels; 50 items elsewhere Total number of levels is 7.
em difficulty	Moderate to easy	Moderate	Moderate to easy	Easy	Moderate to difficult at first two levels; easy a other levels.
rcentage of students scoring tween chance level and 80% rrect on functional level test op figure) and one level above ottom figure).	70.9% 82.0%	83.4% 82.3%	68.5% 81.0%	50.9% 74.5%	57.1% 74.3%
lationship between raw and grade ivalent scores:					
) Range across test levels of e percentage of items where a score change of one point sults in a grade-equivalent ore change of half a year or re.	5.5% - 20.0% (median = 12.5%)	0.0% - 4.5% (median = 0.0%)	4.0% - 11.7% (median = 10.0%)	11.4% - 34.6% (median = 16.0%)	0.0% - 12.0% (median = 0.0%)
) Range across levels of the ximum grade-equivalent change sed by a one-point change in score.	0.5 - 1.5 years (median = 1.2)	0.3 - 0.6 years (median = .4)	0.7 - 2.5 years (median = 1.6)	0.6 - 1.7 years (median = .9)	0.1 - 0.9 years (median = .4)
dian reliabilities at functional vel (top figure) and one level ve (bottom figure).	.80 - .81 .79 - .86	.80 - .88 .76 - .92	not available	.68 - .79 .81 - .82	.84 - .86 .83 - .87

Conclusions

Altogether, the test selection process took two years. In retrospect, a number of things should have been done differently.

First of all, the political (non-measurement) context of the test selection process needs to be determined and taken into account. Is there a hidden agenda which favors staying with a new version of the currently used test? Is there the feeling that a totally new testing series ought to be adopted? Do minority groups feel (with or without basis) that some particular tests are biased? Is there sentiment in favor of adopting an "easier" test or is there horror at the prospect of having to explain to the public tests scores that take a sudden leap? Do people want to use the same test as surrounding communities (or other similar school systems)? Is there concern about whether the local school system is included in a test's norming population? Do people perceive (again rightly or wrongly) some tests to be more in line with local curriculum than other tests? All of these considerations need to be taken into account by the people in charge of test selection.

One way to defuse some of these potential problems is to develop an objective set of criteria for rating the tests. This could include some of the measures of validity, reliability and scale quality that were described in this paper. It is not necessary to conduct a two-year study with testing of large numbers of students at each grade; much can be learned by a systematic examination of test manuals (including technical manuals) and the tests themselves. From the test manuals it is possible to determine test reliabilities; the percent of students in the national population who scored at the floor or ceiling of the test; the relationship between raw

score and grade-equivalent or percentile (does a one-point increase in raw score result in a very large change in G.E. or percentile?); test length (are the tests long enough to adequately sample subject area content); item difficulty (is there a wide range of difficulties? Does the test provide adequate discrimination for both very high- and low-scoring students or do students tend to either "top out" or "bottom out?"); and information on test bias (have meaningful, well-designed bias studies been conducted?)

Similarly, a careful examination of the test can yield a lot of useful information. What does the test itself look like? Is the print legible? Are the graphics clear and unambiguous? Are the answer sheets legible? Do the test items provide sufficiently many distractors or are there many items with only two or three possible alternative answers? Can reading test questions be answered without reading the passages to which they refer?

Finally, what does the test measure? Is it in line with local curriculum? Should it be? One can argue that a set of reading skills tied to a local curriculum does not constitute a more valid definition of "reading" than does the test developers' objectives culled from a wide variety of nationally used basal readers. Still, this is an issue that may need to be addressed.

The process of test selection need not be painful or burdensome. The best way to proceed is to determine at the start exactly what you and your colleagues are looking for in a new test.

References

American Psychological Association. Standards for Educational and Psychological Tests. Washington, D.C.: American Psychological Association, 1974.

Chicago Sun-Times. Easy way out on test scores. January 3, 1982.